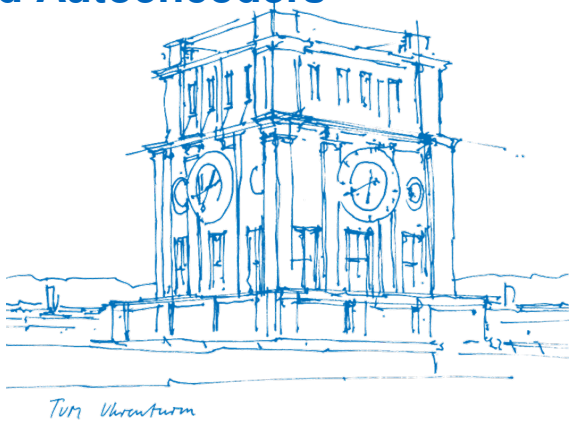
School of Computation, Information and Technology
Technical University of Munich

Associative Memory in Iterated Overparameterized Sigmoid Autoencoders



Omar Faig Orujlu

School of Computation, Information and Technology
Technical University of Munich

August 9th, 2023

Outline

Introduction

- Overparameterized autoencoders can be trained to implement associative memory via iterative maps, when the trained input and output Jacobian of the network has all of its eigenvalues below 1.

Introduction

- Overparameterized autoencoders can be trained to implement associative memory via iterative maps, when the trained input and output Jacobian of the network has all of its eigenvalues below 1.
- In this work mainly sigmoid autoencoders and single and multiple training examples are explored

Introduction

- Overparameterized autoencoders can be trained to implement associative memory via iterative maps, when the trained input and output Jacobian of the network has all of its eigenvalues below 1.
- In this work mainly sigmoid autoencoders and single and multiple training examples are explored
- *Definition 1: A fixed point x^* of map $f(f(x*) = x^*)$ is an attractor if there exists an open neighborhood of x^* such that for any x in this neighborhood , $\{f^k(x)\}_{k\in N}$ converges to x^* as $k \to \infty$. The set of all such points is called basin of attraction of x^* .*

Introduction

- Overparameterized autoencoders can be trained to implement associative memory via iterative maps, when the trained input and output Jacobian of the network has all of its eigenvalues below 1.
- In this work mainly sigmoid autoencoders and single and multiple training examples are explored
- *Definition 1: A fixed point x^* of map $f(f(x*) = x^*)$ is an attractor if there exists an open neighborhood of $x*$ such that for any x in this neighborhood , $\{f^k(x)\}_{k \in N}$ converges to x^* as $k \to \infty$. The set of all such points is called basin of attraction of x^* .*
- *Proposition 1: A fixed point x^* is an attractor of a differentiable map f if all eigenvalues of the Jacobian of f at x^* are strictly less than 1 in absolute value*

Introduction

- Therefore, for a point x to be an attractor following conditions should be fulfilled: (1) $f(x) = x$ and (2) all the eigenvalues of the $J(x)$ have norm strictly smaller than 1.
- First conditioned can be theoretically achieved in NTK limit if all data points are sampled from unit sphere, the network has non-polynomial Lipschitz activation function and $L \geq 2$
- In practice it is easy to fulfill the first condition that's why in paper main focus was on the second condition.

Introduction

- The Jacobian matrix is calculated using following formula :

$$J(x) = \frac{1}{\sqrt{n_L}} W^{(L)} \prod_{k=1}^{L} \left(D^{(k)} \frac{1}{\sqrt{n_{k-1}}} W^{(k-1)} \right)$$

where

$$D^{(k)} = diag(\dot{\sigma}(\alpha^{(k)}(x))) \tag{1}$$

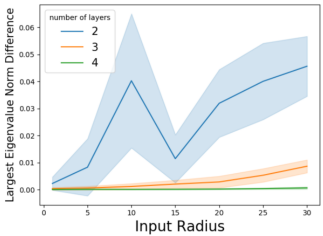
- The following loss function is used:

$$argmin_f \frac{1}{2n} \sum_{i=1}^{n} \|f(x_i) - x_i\|_2^2 \tag{2}$$

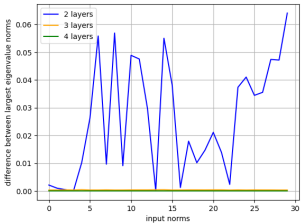
- The threshold for training loss is 10^{-7}.The threshold for iterative convergence is 10^{-2}.

Outline

Single Training example



(a) Results from paper



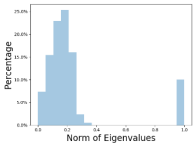
(b) Reproduced result

The results from single training example are provided above. Here sigmoid network has hidden dimension 1000 and input dimension 32. Here it can be observed that the difference between largest eigenvalues of input and output Jacobians decrease with more number of layers.

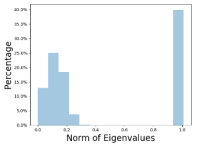
Multiple Training Examples
Linear Region

- In the paper it is shown that for small input norm r , a network in the NTK limit behaves like a linear network and eigenvalues of Jacobians are 1 and due to that attractor formation can fail.
- For illustration purposes the unit vector is sampled as data. The input dimension of sigmoid network is 10 and hidden size is 1000. 2,5 and 8 points are trained and as suggested by Lemma 4 in the paper , there should be $n-1$ eigenvalues with norm around 1.

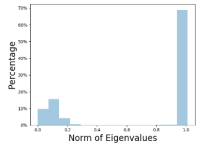
Multiple Training Examples
Linear Region



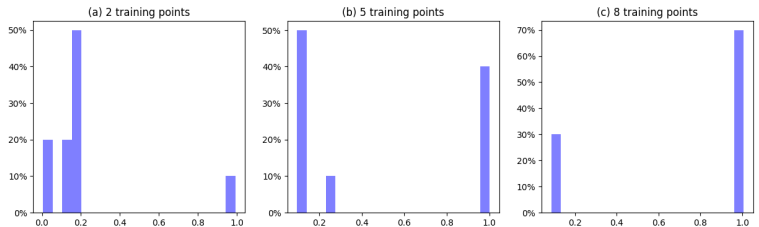
(a) 2 training points



(b) 5 training points



(c) 8 training points

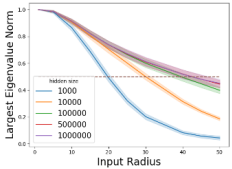


Eigenvalue distribution of 2 - layer sigmoid network trained with input dimnesion 10

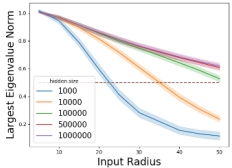
Multiple Training Examples
Beyond Linear Region

- In order to get the network beyond linear region the input dimension is chosen 32. Here different hidden sizes and 5,20 and 40 training points are used
- The reproducibility results here are sligtly different, namely, only 2,5 and 10 training points with hidden size 1000 and 10000 are used. Nevertheles, similar tendency can be observed as in the paper.

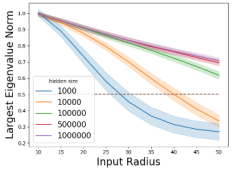
Multiple Training Examples
Beyond Linear Region



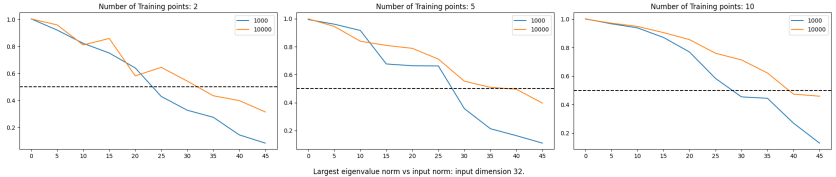
(a) number of training points: 5



(b) number of training points: 20

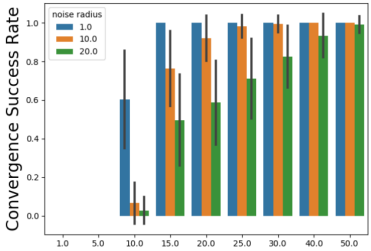


(c) number of training points: 40

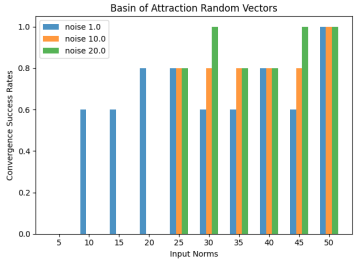


Basin of attraction

- For testing basin of attraction Gaussian noise is added to the input values and checked whether the network can converge to the original input under 50 iterations. Here the input dimension is 32 and hidden layer size is 10000 with 2 layers. Here it can also be observed that the larger input norms give bigger basin of attraction.



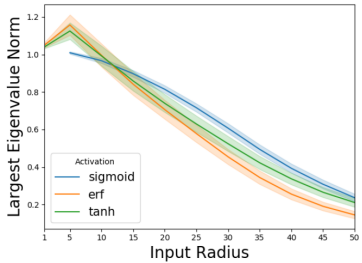
(a) Results from paper



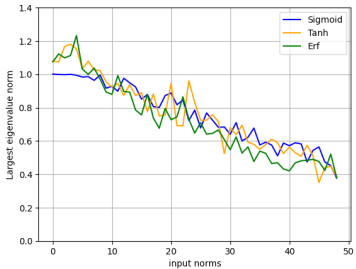
(b) Reproduced result

Different Activation Functions

- As stated in paper the results can be extended to different activation functions. For this following graph is generated, where 2 layer networks with hidden size 10000 and input dimension 32 are trained.



(a) Results from paper



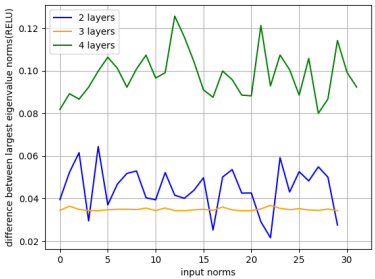
(b) Reproduced result

Outline

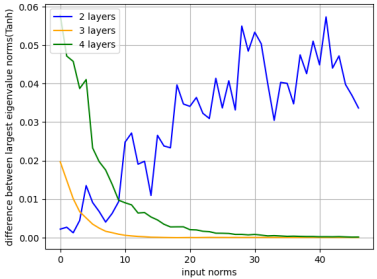
Extension 1

Lipschitz and Non-Lipschitz activation functions

- For testing the importance of Lipschitz activation function first experiment is repeated with Tanh and ReLU activation fucntions



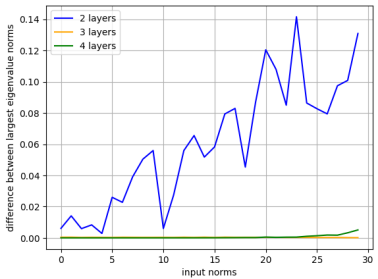
(a) ReLU



(b) Tanh

Extension 2
NTK Limit

- In order to show the importance of NTK limit the first experiment from the paper is repeated with hidden size of 50 instead of 1000. In the diagramm it can be seen that the diference between largest eigenvalues of Jacobians in 2 layer case is higher than the paper.



(a) Results with hidden size 50

Extension 3

Attractor vs Generalization

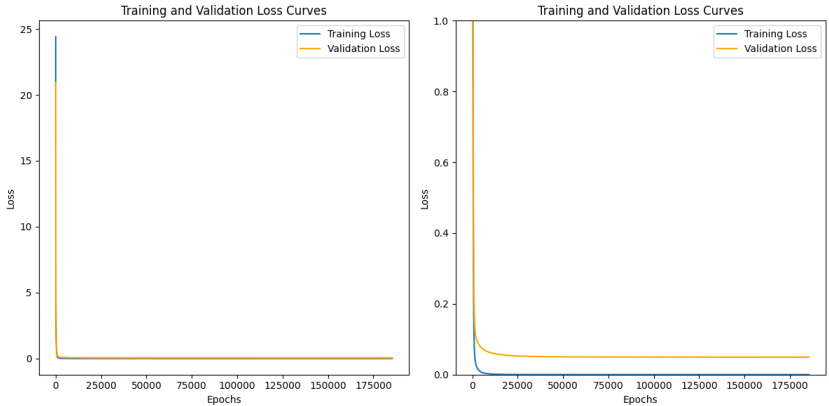
- In order analyze the relationship between attractor formation and generalization of deep learning multiple experiments are conducted. In the following table the important setup parameters are shown:

Input Norm	Hidden Size	Training Points	Validation Points
15	1000	100	30
15	1000	50	20
15	100	100	30
15	50	100	30
mixed	1000	100	30

Table 1 Experimental Setup

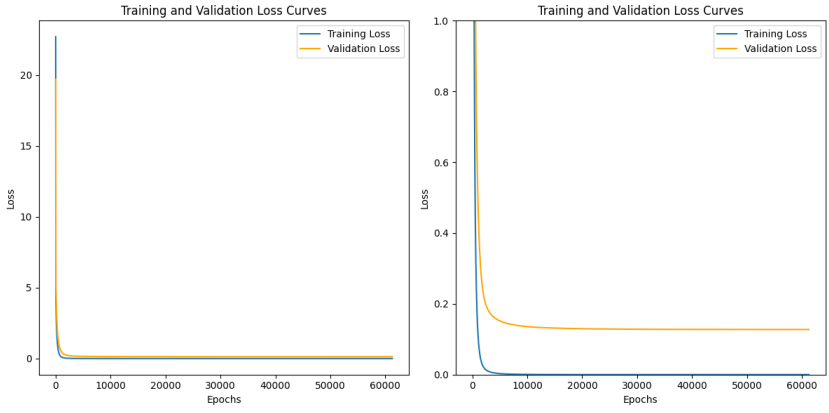
Extension 3
Attractor vs Generalization

- Hidden size 1000, norm radius 15 and 100 training points



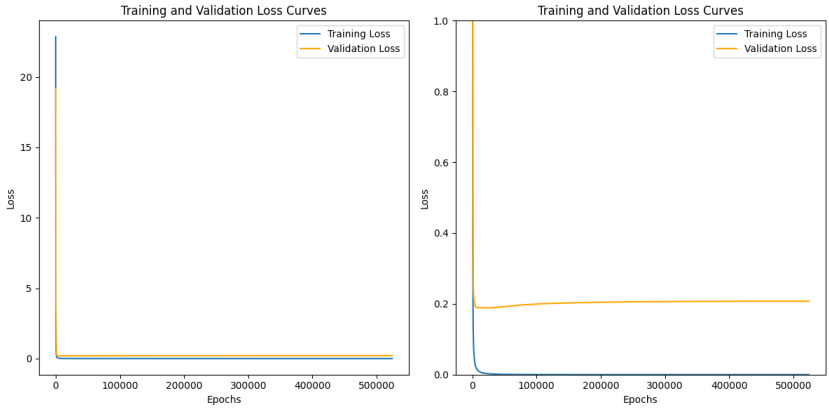
Extension 3
Attractor vs Generalization

- Hidden size 1000, norm radius 15 and 50 training points



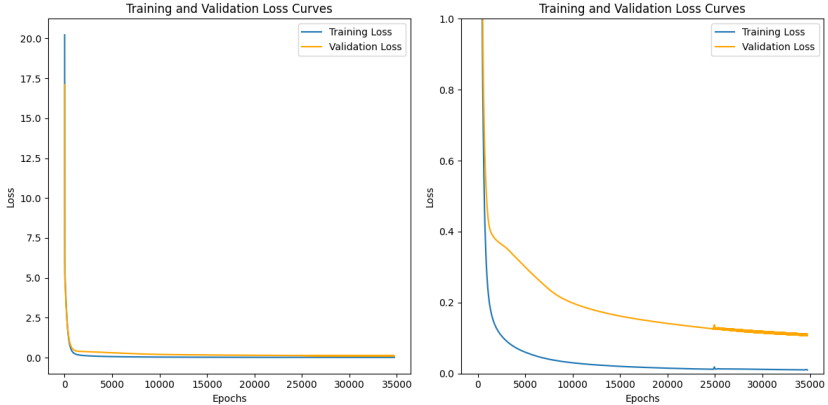
Extension 3
Attractor vs Generalization

- Hidden size 100, norm radius 15 and 100 training points



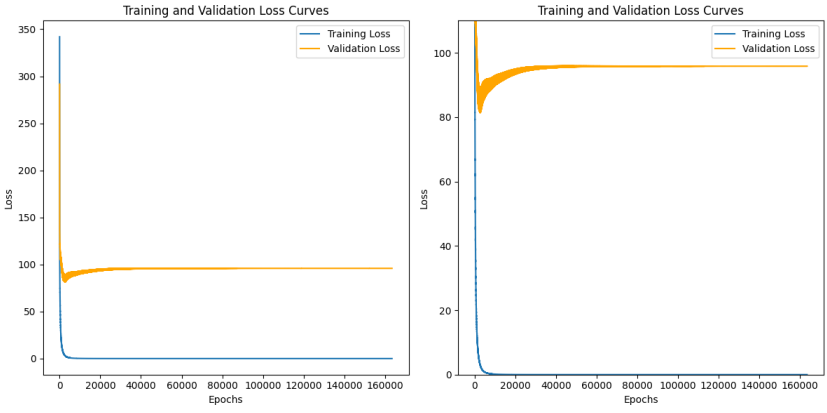
Extension 3
Attractor vs Generalization

- Hidden size 50, norm radius 15 and 100 training points and loss threshold is 10^{-2}(Non-attractor)



Extension 3
Attractor vs Generalization

- Hidden size 1000, norm radius Mixed and 100 training points



Resources

- Jiang, Y., Pehlevan, C. Associative Memory in Iterated Overparameterized Sigmoid Autoencoders. International conference on machine learning, 2020